iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: https://doi.org/10.22214/ijraset.2026.85490

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue X Oct 2026- Available at www.ijraset.com

AttritionSense AI: An Intelligent Employee Attrition Prediction and Retention Recommendation System Using Random Forest

Anitha D[1], Anitha E[2], Divya Dharshini V[3]
Department of CSE, [Adiyamaan College Of Engineering], [Hosur], India

Abstract: ***Employee attrition is costly for organisations, yet HR teams usually learn that someone is leaving only after the resignation. This paper presents AttritionSense AI, a role-based web application that predicts the attrition risk of an employee, explains the likely drivers and recommends concrete retention actions. A Random Forest classifier, trained on the 1,470-record IBM HR Analytics dataset with class-weight balancing, converts the predicted probability into a 0–100 risk score and a Low, Medium or High tier. A transparent rule layer lists the specific risk drivers of each employee, derives a growth-status indicator and raises a Priority Retention alert only when an employee of high growth potential is also at high risk. Prescriptive action plans with an owner and an urgency target are generated for HR, and an executive PDF report can be exported. A separate, private employee portal offers a conversational companion (BuddyBot), a weekly well-being quiz, relaxation games and an eight-dimension feedback form that are deliberately kept out of the risk score. On a stratified 20% hold-out set the model reaches an ROC-AUC of 0.798 and an attrition recall of 0.489 at a precision of 0.479; accuracy is 0.833. We discuss why accuracy alone is misleading for this imbalanced problem and list the limitations and the planned extensions.***

Keywords: ***employee attrition, random forest, retention recommendation, decision support, role-based access, Streamlit, scikit-learn, people analytics***

I. INTRODUCTION

Losing an employee means paying for recruitment and training again, losing knowledge and disturbing the team. The difficulty for HR is timing: the signs of dissatisfaction, such as constant overtime, a stalled career path or a poor work–life balance, are present in the records long before the resignation, but they are rarely read together. Machine-learning models trained on historical HR data can combine such signals into a single early-warning score [1].

A score alone is not enough for a manager. HR also needs to know why a person is at risk and what to do about it, and employees need to trust that a prediction tool will not be used against them. This paper describes AttritionSense AI, a web application built around these three needs: prediction, explanation and action, with a strict separation between what HR can see and what employees keep private. The main contributions are:

1) *Automatic risk prediction:* HR selects a verified employee number and the system loads all attributes from the dataset, so no model input is typed by hand. A Random Forest returns an attrition probability, a 0–100 risk score and a Low, Medium or High tier.
2) *Explanation and prescription:* Ten rule-based risk drivers, a growth-status indicator and a Priority Retention rule turn the score into a tailored action plan with an owner and a deadline, exportable as a PDF dossier.
3) *Privacy-preserving well-being support:* A private employee portal with a chatbot, quiz, games and feedback form is kept strictly decoupled from the risk score and from performance reviews.

The rest of the paper is organised as follows. Section II reviews related work, Section III states the problem, Section IV presents the architecture, Section V explains the methodology, Section VI reports the implementation and results, Section VII discusses limitations and Section VIII concludes.

II. LITERATURE REVIEW

Attrition prediction is usually treated as binary classification on tabular HR records. The IBM HR Analytics Employee Attrition dataset [1] is a widely used public benchmark for this task: it holds 1,470 employee records with 35 attributes covering demographics, job role, income, satisfaction ratings, overtime and tenure.

©IJRASET: All Rights are Reserved

Random forests [2] are a natural choice for such data because they combine many decorrelated decision trees, handle numerical and categorical attributes together, capture non-linear interactions and expose feature importances. They are available in scikit-learn [3]. Gradient-boosted trees such as XGBoost [4] are a common alternative.

Attrition datasets are imbalanced, because most employees stay. Two common remedies are synthetic oversampling such as SMOTE [5] and cost-sensitive learning through class weights; we use the latter, which leaves the data unchanged. Many published attrition models stop at a probability. AttritionSense AI differs in that it adds human-readable drivers, a growth-aware retention rule, prescriptive actions and a confidential support channel for employees, so that the output can be acted upon rather than only read.

III. PROBLEM STATEMENT AND OBJECTIVES

Given the verified record of an employee, the system must (a) estimate the probability that the employee will leave, (b) express it as an understandable risk tier, (c) identify the concrete factors behind the risk, (d) recommend retention actions with an owner and an urgency, and (e) give employees a private space for well-being support that is never used to score them. The objectives are to make prediction fully automatic, to keep the logic transparent and testable, and to enforce role separation so that HR cannot read private chats and employees cannot see organisational analytics.

IV. SYSTEM ARCHITECTURE

A. Overview

The application follows a three-tier design (Fig. 1). A Streamlit front end provides separate employee and HR portals. The application tier is a set of Python modules for authentication, prediction, recommendation, growth analysis, chatbot, e-mail and reporting. The data tier consists of a SQLite database, the employee CSV file and the serialised model.

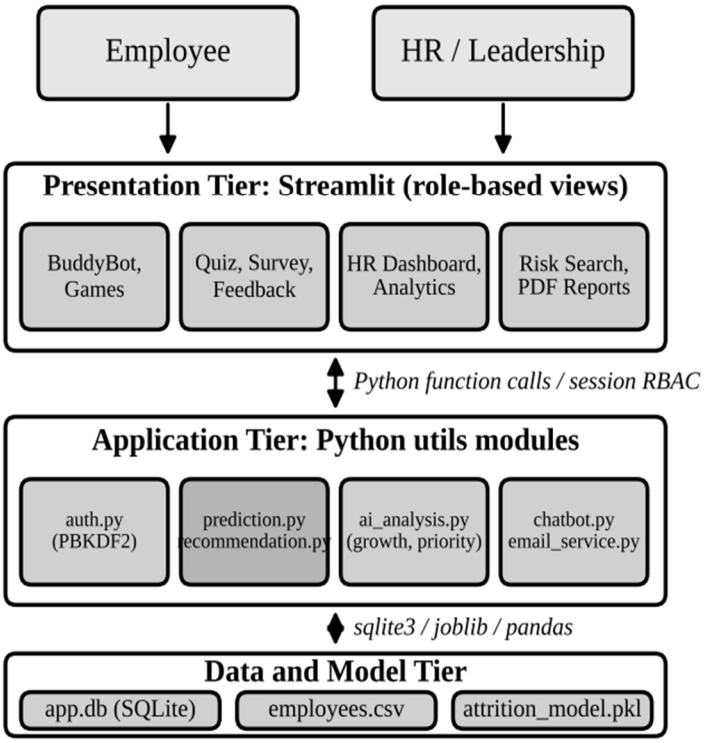


Fig. 1 Three-tier architecture of AttritionSense AI

B. Technology Stack

Table I lists the technologies used. All are open source and widely documented, which keeps deployment simple.

TABLE I
TECHNOLOGY STACK

Layer	Technology
Interface	Streamlit [6], Plotly [8] charts
Machine learning	scikit-learn [3] Random Forest, Joblib
Data handling	Pandas, NumPy, CSV files
Storage	SQLite [7] (app.db)
Security	PBKDF2-HMAC-SHA256 with per-user salt [10]
Reporting	ReportLab [9] PDF dossiers
Notification	Python smtplib e-mail alerts
Chatbot	Rule-based engine; optional Gemini API [11]

C. Database Design

The SQLite schema has eight tables: users (identifier, role, password hash and salt), chat_history, feedback (eight 1–5 ratings, anonymity flag and comments), prediction_history (probability, risk score, tier, drivers, recommendations, growth status, confidence and priority flag), surveys, quiz_scores, exit_feedback and retention_alerts. Prediction and feedback records are also mirrored to CSV files for export.

D. Role-Based Access

Two roles are supported. Every page is guarded by a session check, and a user who is not logged in under the correct role is redirected to the matching login page. Table II summarises the permitted functions. Employee accounts are seeded only for employee numbers that exist in the dataset.

TABLE II
ROLES AND PERMITTED FUNCTIONS

Role	Permitted functions
Employee	Read-only profile, BuddyBot chat, games, weekly well-being quiz, survey and 8-dimension feedback
HR / Leadership	Executive dashboard, employee search and prediction, workforce analytics, feedback analytics, history, PDF reports

V. METHODOLOGY

A. Dataset and Preprocessing

We use the 1,470-record IBM HR dataset [1] with 35 attributes, of which 237 employees (16.1%) have left. The identifiers EmployeeNumber, EmployeeCount, Over18 and StandardHours are dropped. The remaining 23 numerical attributes are standardised and the 7 categorical attributes (BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime) are one-hot encoded inside a single scikit-learn pipeline, so the same transformation is applied at training and at prediction time. The data are split 80/20 with stratification (1,176 training and 294 test records, random seed 42).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Random Forest Model

The classifier is a Random Forest [2] with 250 trees, maximum depth 12, a minimum of 6 samples to split, a minimum of 2 samples per leaf and balanced class weights, which raise the penalty for misclassifying the minority class of leavers.

C. Risk Score and Tiers

For an employee with feature vector x, the model outputs the probability $p = P(\text{leave} \mid x)$. The risk score is

$$R = 100 \times p \quad (1)$$

and is rounded to one decimal. Scores of 0–39 are labelled Low Risk, 40–69 Medium Risk and 70–100 High Risk. The prediction confidence is $\max(p, 1 - p) \times 100\%$. Note that the score is the raw forest probability and has not been separately calibrated.

D. Risk Driver Identification

Because a forest is hard to read directly, ten transparent rules examine the employee's own attributes and report concrete drivers with a severity: frequent overtime; job, environment or work–life-balance ratings of 2 or less; four or more years since the last promotion (with at least three years in the company); a commute of 15 miles or more; no stock options; frequent business travel; less than a year with the current manager; and a salary hike of 12% or less. These rules explain the situation of the person; they are not a decomposition of the model's output.

E. Growth Status and Priority Retention

A composite growth score from 0 to 100 is built from the components in Table III, with an optional ±5 adjustment from submitted feedback. Missing feedback is never penalised, and game, quiz and chat activity is never used. A score of 68 or more gives High Growth Potential, a score below 42 (or a performance rating of 2 or less) gives Needs Support, and everything else is Stable Contributor. A Priority Retention alert is raised only for the single combination High Growth Potential and High Risk; the alert is stored once per employee and an HR e-mail is sent if SMTP is configured.

TABLE III
COMPONENTS OF THE GROWTH SCORE

Component	Max pts	Rule (points)
Performance rating	35	≥4: 35; 3: 25; else 10
Job involvement	25	4: 25; 3: 20; 2: 10; 1: 0
Training last year	15	≥3: 15; 2: 10; 1: 5
Salary hike	15	≥18%: 15; ≥14%: 10; ≥12%: 5
Promotion pace	10	≤1 yr: 10; ≤3 yr: 5; ≥7 yr: −5

F. Retention Recommendations

Each detected driver is mapped to an action with a category, an owner and an urgency: overtime or work–life-balance issues lead to an overtime cap and flexible scheduling (immediate, within 7 days); promotion stagnation or a long time in the same role leads to a career-planning session (within 14 days); low satisfaction leads to a role re-alignment interview (within 14 days); and environment, compensation and commute or travel issues lead to team-climate, benchmarking and flexibility reviews (within 30 days).

G. Employee Well-Being Modules

The employee portal contains BuddyBot, a multi-turn companion that classifies the intent of each message (for example workload, meetings, exhaustion or career) and answers from a rule-based engine, optionally using the Gemini API [11]. A crisis-keyword check returns a supportive response, and chats are visible only to the employee. A weekly quiz of 15 non-medical questions is scored out of 75 into a well-being percentage and indicators for stress, work–life balance and career motivation. The eight-dimension feedback form can be anonymous and notifies HR by e-mail. None of these data enters the attrition model or the growth score.

VI. IMPLEMENTATION AND RESULTS

A. *User Interface*

The HR portal has a dashboard with KPI cards and a priority-intervention queue, an employee search with prediction and action plan, Plotly analytics by department, job role, overtime, satisfaction, income and tenure, feedback analytics, a searchable prediction history with CSV export and a report centre (Table IV). The test folder contains automated tests for authentication, prediction and PDF generation, the chatbot, the feedback and e-mail workflow and the HR analytics.

TABLE IV
MAIN APPLICATION MODULES

Module	Purpose
auth.py	PBKDF2 hashing, login, session roles
ml_pipeline.py	Preprocessing, training, evaluation
prediction.py	Inference, risk tier, risk drivers
ai_analysis.py	Growth status, priority retention
recommendation.py	Action plan with owner, urgency
chatbot.py	BuddyBot intent engine
quiz.py, survey.py	Well-being quiz and survey
pdf_report.py	Executive PDF dossier
email_service.py	SMTP alerts to HR

B. *Model Evaluation*

Table V gives the per-class results on the 294 test employees, derived from the confusion matrix [[222, 25], [24, 23]]. Overall accuracy is 0.833 and ROC-AUC is 0.798. Of the 47 employees who actually left, 23 were identified (recall 0.489); of the 48 employees flagged as leavers, 23 were correct (precision 0.479).

TABLE V
PER-CLASS RESULTS ON THE HOLD-OUT TEST SET

Class	Precision	Recall	F1	Support
Stays	0.902	0.899	0.900	247
Leaves	0.479	0.489	0.484	47
Accuracy			0.833	294

These figures must be read carefully. Since 247 of the 294 test employees (84.0%) stayed, a model that always predicts “stays” would reach a higher accuracy than ours; the useful measures are therefore recall on the leaver class and ROC-AUC. The model finds about half of the leavers at the cost of roughly one false alarm per true detection, which is acceptable for a screening tool whose output is reviewed by HR but not for automatic decisions.

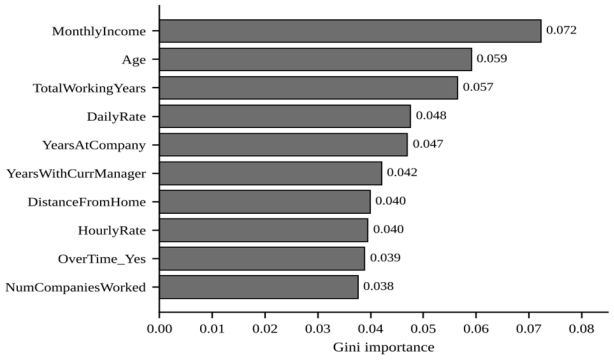


Fig. 2 Ten most important features of the Random Forest model

Fig. 2 shows the ten most important features. Monthly income, age, total working years, daily rate and years at the company lead the ranking, followed by tenure with the current manager, distance from home, overtime and the number of companies worked for. The raw data agree with the role of overtime: 30.5% of employees who work overtime left, against 10.4% of those who do not. Attrition also varies strongly by job role, from 2.5% for research directors to 39.8% for sales representatives.

VII. LIMITATIONS AND FUTURE WORK

The current system is a working prototype and has limitations that we plan to address.

1) *Data:* The IBM dataset is a public, fictional benchmark. The model should be retrained and validated on an organisation's own records before use.
2) *Evaluation:* Results come from a single stratified split. Cross-validation, hyper-parameter search, threshold tuning and comparison with boosted trees [4] or SMOTE [5] are needed, and the probabilities should be calibrated.
3) *Explainability:* The risk drivers are rules, not model attributions. Per-employee explanations such as SHAP values would tie the explanation to the prediction.
4) *Fairness:* Age, gender and marital status are model inputs. A fairness audit, or removing such attributes, is required before real deployment.
5) *Security:* The demonstration accounts use default passwords, which must be replaced, and secrets should be kept in environment variables or a secret store.

Further work includes a live HR-system integration, outcome tracking to learn which retention actions work, and a user study with HR practitioners.

VIII. CONCLUSIONS

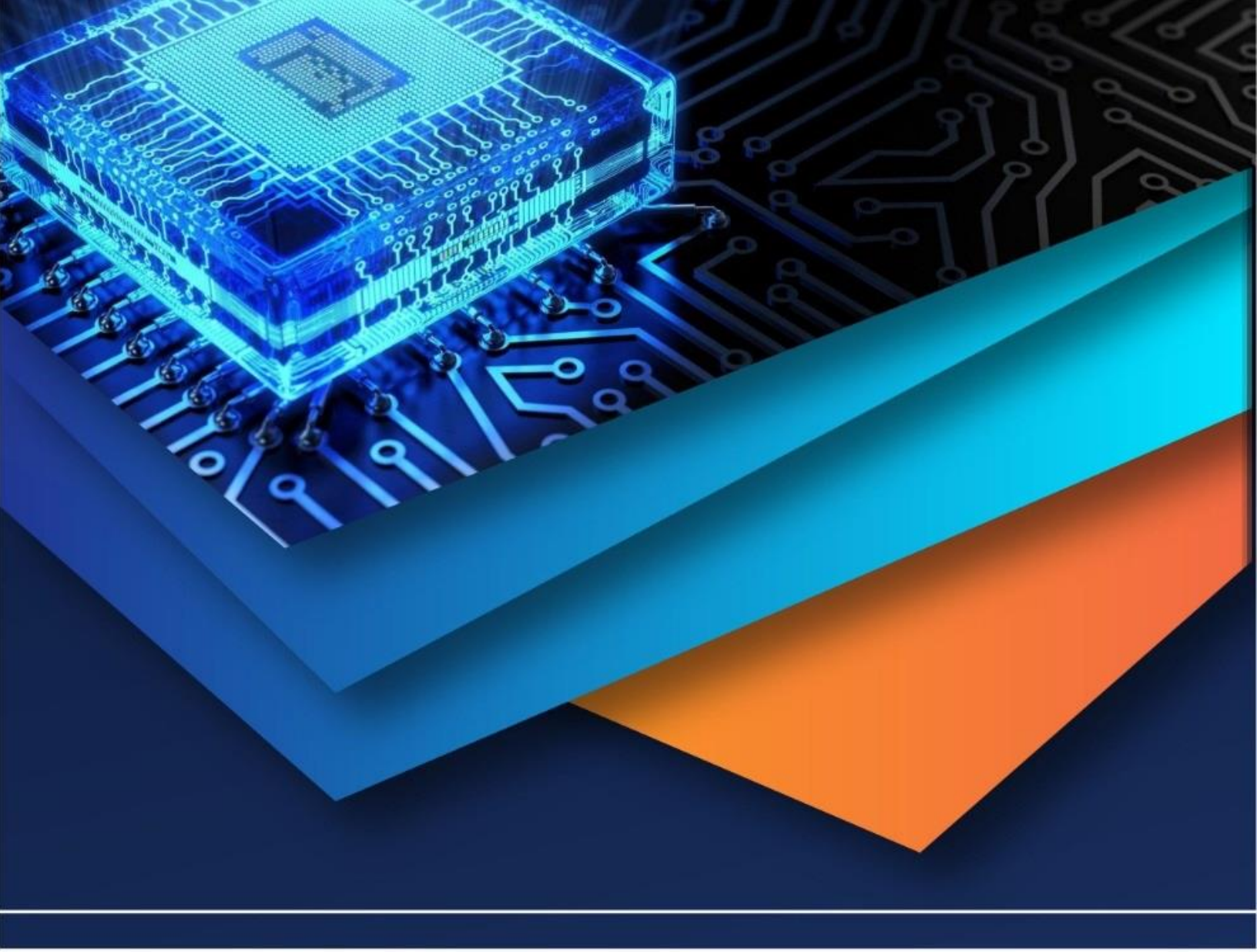
We presented AttritionSense AI, a role-based system that combines automatic Random Forest attrition prediction, transparent risk drivers, a growth-aware Priority Retention rule, prescriptive action plans and PDF reporting with a private well-being portal for employees. On the IBM benchmark the model reaches an ROC-AUC of 0.798 and detects 49% of leavers, which supports its use as a screening and decision-support aid under human review. Validation on real organisational data and a fairness audit are the next steps.

IX. ACKNOWLEDGMENT

The authors thank V.Anbu Priyadharshini for guidance and support during this work.

REFERENCES

[1] IBM, "IBM HR Analytics Employee Attrition & Performance," Kaggle dataset. [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
[2] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[3] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
[5] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[6] Streamlit Inc., "Streamlit documentation." [Online]. Available: https://docs.streamlit.io/
[7] SQLite Consortium, "SQLite documentation." [Online]. Available: https://www.sqlite.org/docs.html
[8] Plotly Technologies Inc., "Plotly Python graphing library." [Online]. Available: https://plotly.com/python/
[9] ReportLab Inc., "ReportLab PDF toolkit." [Online]. Available: https://www.reportlab.com/
[10] M. S. Turan *et al., Recommendation for Password-Based Key Derivation, Part 1: Storage Applications*, NIST SP 800-132, 2010.
[11] Google, "Gemini API documentation." [Online]. Available: https://ai.google.dev/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)